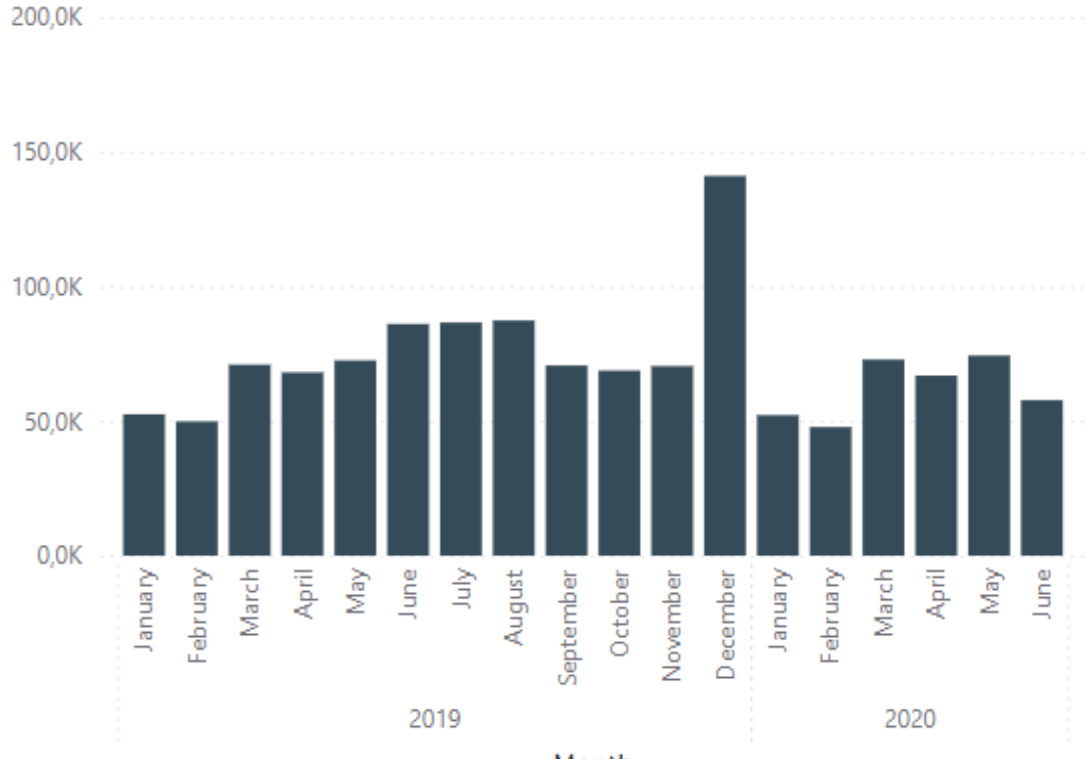
Credit Card Purchases & Fraud Detection | By Liam Jameson

Data Overview | Data Exploration | Regression Model

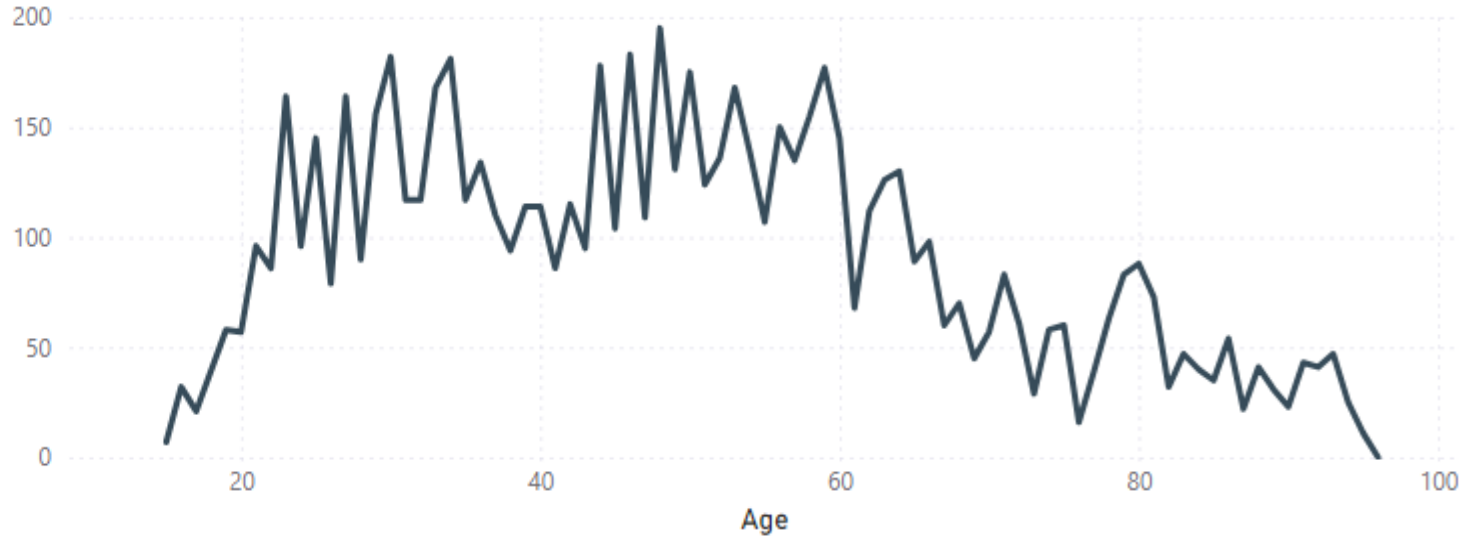
Random Sample of Data

amt	city	city_pop	gender	dob_year	dob_month	dob_day	is_fraud	job	Year	Quarter	Month	Day	long	lat	state	merchant
65,79	Birmingham	888	F	1988	3	25	0	Camera operator	2019	Qtr 1	March	9	-91,95	40,86	IA	fraud_Sporer Inc
54,34	Bonfield	1617	F	1990	4	25	0	Medical secretary	2019	Qtr 4	November	27	-88,06	41,16	IL	fraud_Dach-Borer
2,83	Detroit	673342	M	1983	9	2	0	Health visitor	2020	Qtr 1	March	16	-82,99	42,37	MI	fraud_Spencer-Runolfsson
1 697,19	Matthews	1019	F	1979	1	26	0	Aeronautical engineer	2020	Qtr 1	February	9	-89,63	36,72	MO	fraud_Keeling-Crist
12,01	Mifflin	1909	F	1954	8	22	0	Mining engineer	2019	Qtr 4	December	28	-77,40	40,56	PA	fraud_Gerhold LLC
9,38	Mount Morris	4895	F	1958	10	29	0	Acupuncturist	2019	Qtr 4	December	9	-77,87	42,68	NY	fraud_Friesen Inc
9,08	Mulberry Grove	1810	F	1974	12	24	0	Race relations officer	2019	Qtr 3	July	21	-89,25	38,93	IL	fraud_Kemmer-Buckridge
9,65	Smackover	2501	M	1986	6	11	0	Financial adviser	2019	Qtr 1	February	13	-92,74	33,34	AR	fraud_Morissette PLC
1,03	Tamaroa	2135	M	1961	1	31	0	Development worker, community	2020	Qtr 1	March	10	-89,22	38,14	IL	fraud_Lind-Buckridge
54,98	Thomas	1675	F	1986	5	1	0	Barrister	2019	Qtr 3	September	12	-98,74	35,74	OK	fraud_Bednar PLC

Value of Fraudulent Transactions by Month



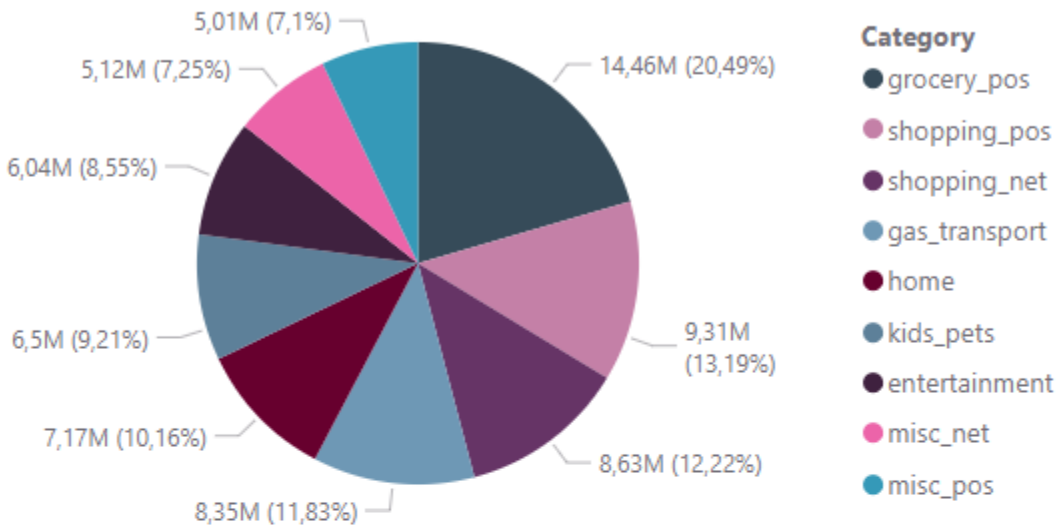
Fraud Cases by Birth Year



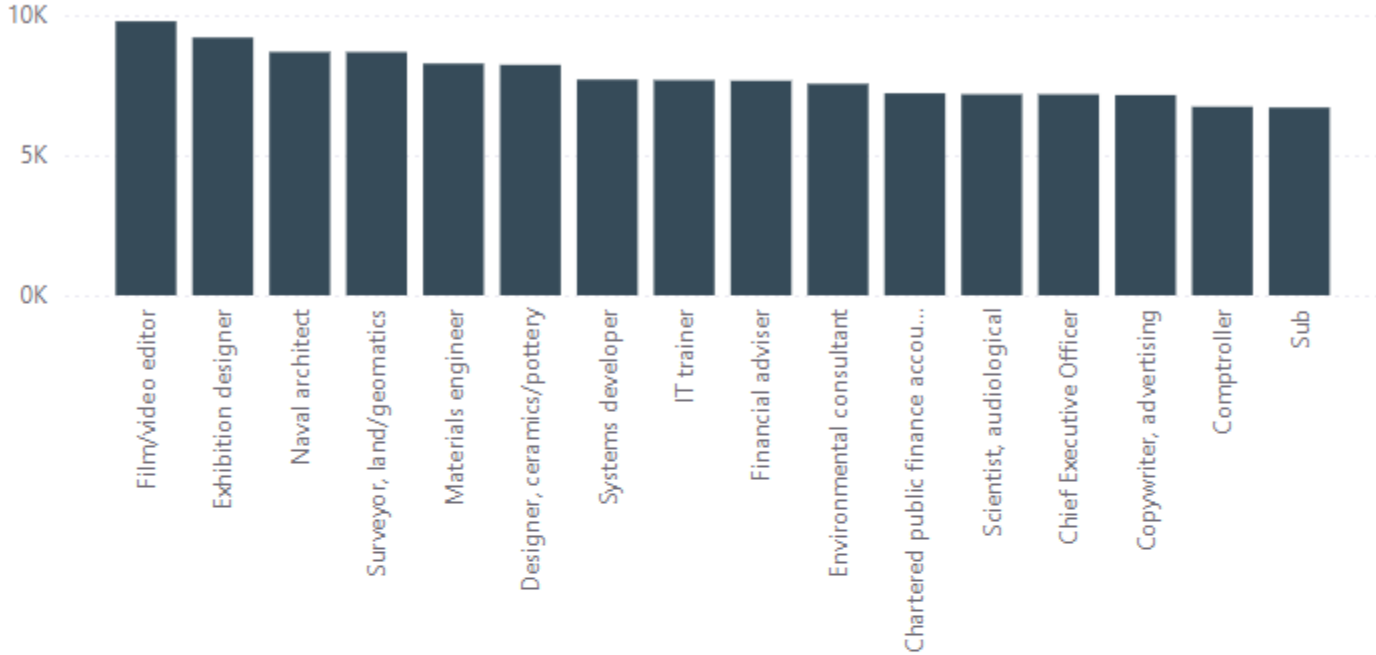
Number of Data Points in the Dataset	Average Value of Transactions	Number of Fraud Cases	Average City Population
1.30M	70.35	7506	88,82K

Credit Card Purchases & Fraud Detection | By Liam Jameson

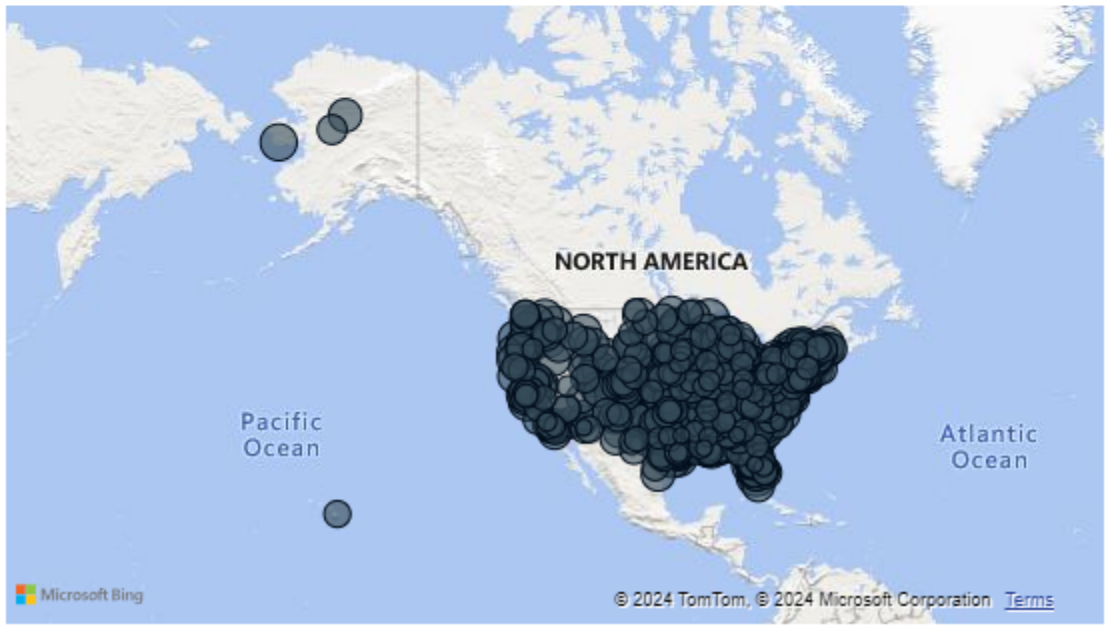
Percentage of Sales by Category



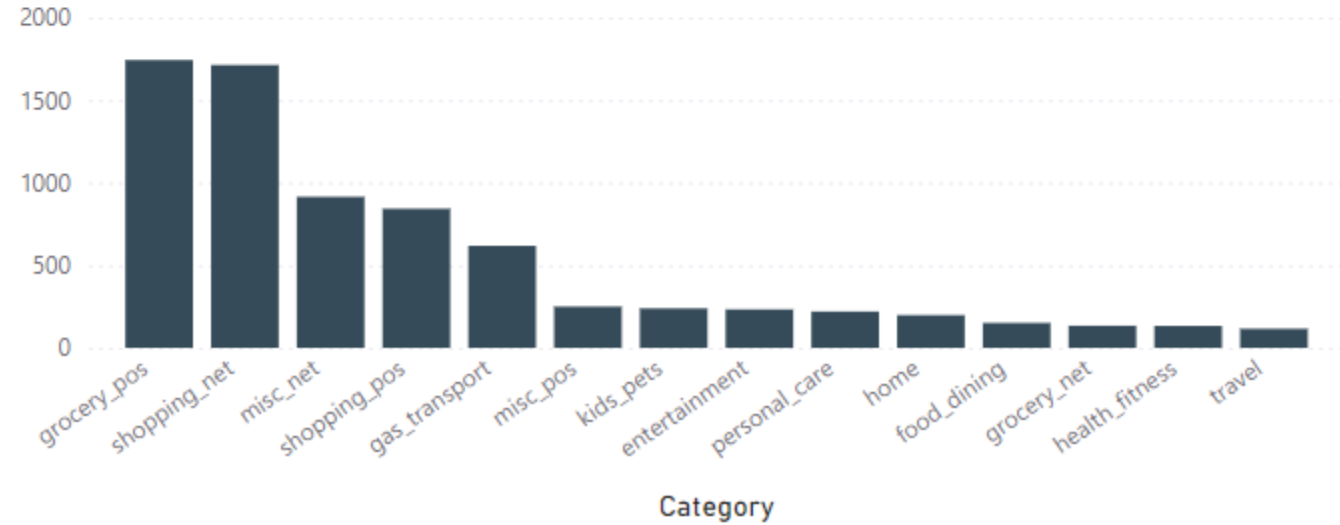
Count of Transactions by Profession



Regional Distribution of Credit Card Purchases

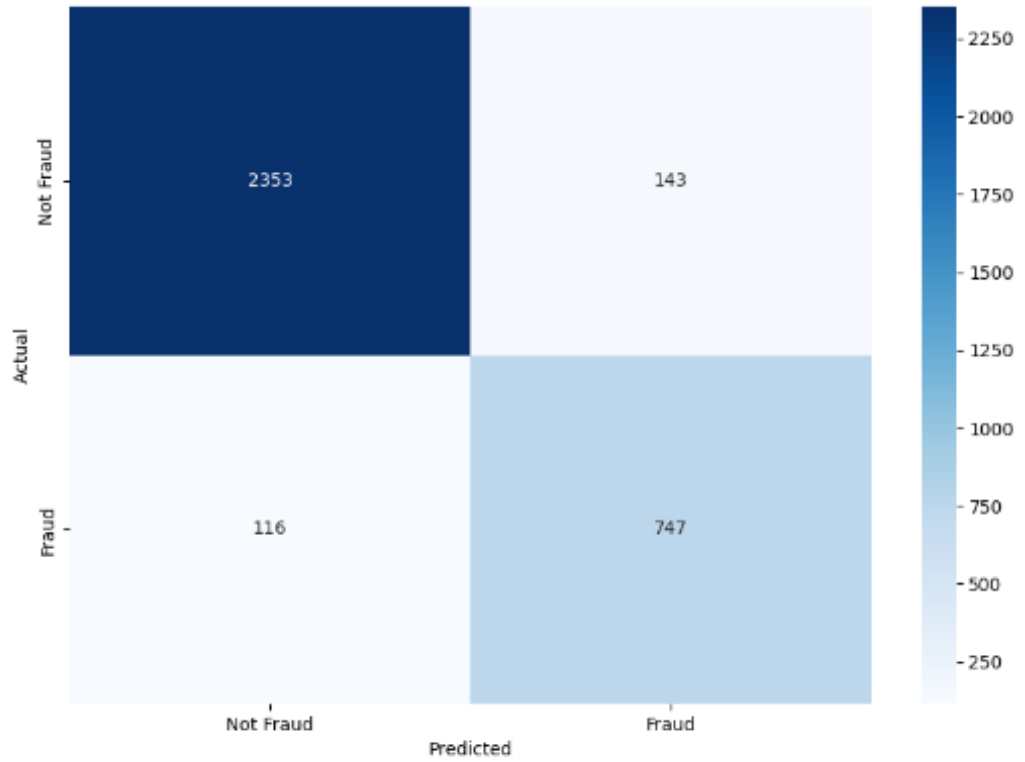


Number of Fraud Cases by Purchase Category



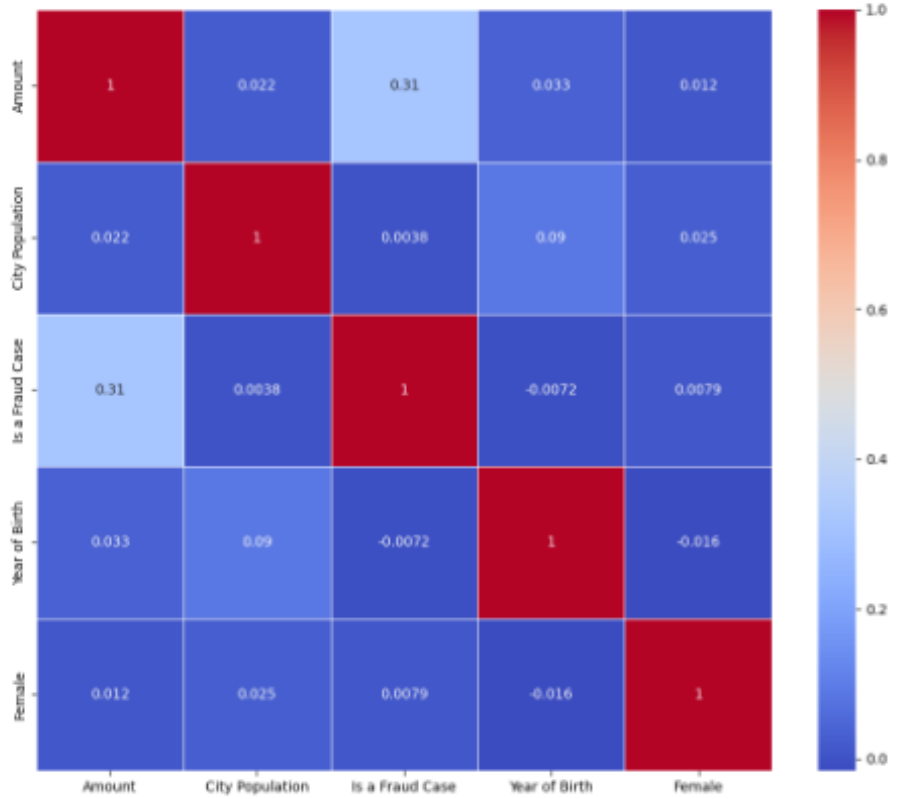
Credit Card Purchases & Fraud Detection | By Liam Jameson

Confusion Matrix for the Regression Model



This is the Confusion Matrix for a regression model that was fitted based on 7 of the variables present in the data set. As we can see from the dataset the model correctly predicts around 92% of cases given to it. This could be improved further through more data becoming available or more normalized job categories or time values in further analysis.

Confusion Matrix for Selected Variables from the Model



This correlation matrix shows us which variables from our data set have the greatest impact on fraud, higher numbers represent a higher correlation with fraud and vice versa. As we can see from the matrix, city population and being female both have positive effects on falling victim to fraud. However, the biggest positive correlation is between the amount of a transaction and fraud.